AUDIONYX: REAL-TIME DETECTION OF AUDIO DEEPFAKES IN PHONE CALLS

Emmanuel J Jose
Dept. of Computer Science
Toc H Institute of
Science and Technology
Kochi, India
emmanueljjose10@gmail.com

Fidha Fathima N S
Dept. of Computer Science
Toc H Institute of
Science and Technology
Kochi, India
fidha1610@gmail.com

Gautham Babu
Dept. of Computer Science
Toc H Institute of
Science and Technology
Kochi, India
gauthambabu12@gmail.com

Liya Latheef
Dept. of Computer Science
Toc H Institute of
Science and Technology
Kochi, India
liyalatheef23@gmail.com

Shanthi N.M
Asst. Professor, Dept. of CS
Toc H Institute of
Science and Technology
Kochi, India
shanthinm@tistcochin.edu.in

Abstract

The explosion of AI-assisted voice synthesis technologies has made audio deepfake–based fraud a greater risk, especially within telecommunication domains. These synthetic voices are one of the leading impersonation methods, attacks and scams with potentially grave security hazards. Detecting real-time deepfakes is challenging due to bandwidth limitations, codec compression, and background noise that obscure distinguishing artifacts. This paper presents Audionyx, a real-time audio deepfake detection framework that is intended for telephony applications. It uses a lightweight custom Convolutional Neural Network (CNN) trained on Mel-spectrogram abstractions to strike an optimal balance between accuracy in detection and computational efficiency. A sliding window segmentation strategy and probabilistic aggregation mechanism ensure stable and reliable detection across continuous audio streams. Experimental evaluation demonstrates excellent detection performance and low latency, testing the ability of the system to be deployed in real time. The proposed approach is a robust and scalable method for reducing fraud through voice and for improving security against impersonation attacks.

Index Terms: Audio deepfakes, real-time detection, Telephony channels, CNN-Transformer, Mel spectrogram, voice fraud detection

I. Introduction

The rapid development of Artificial Intelligence (AI) has delivered super-sophisticated generative models able to model synthetic media, popularly referred to as deepfakes. They can produce complex data sets and even generate artificial art. But, these technologies are not without good intentions and good reasons, they also have serious threats, especially in audio. The technology for generating new solutions for audio and video editing. Cloning someone's voice in this study TTS can clone a person using Text-to-Speech (TTS) technology: using Text-to-Speech Technology is capable of imitating the output as in-mute a person's voice from a small audio sample using TTS and voice-transformation technologies are available for everyone to the masses and voice conversion technologies. Democratization of powerful AI tools that has made empowerment democratization of high AI tools has allowed bad actors to launch a new level of serious that involve more and sophisticated phone scams that can be an entirely new order fake synthetic voice that increasingly rely on synthetic voices to commit sophisticated phone scams—of the new generation of sophisticated phone frauds with synthetic sounding, that rely on synthetic by the process, bank officials employees—the purpose is to make a little-known financial fraud, misinformation, and that leads to a lot of personal distress. The communication network also has its own limitations with its own set of special capabilities and restrictions. An environment of difficulty in detecting manipulations of this kind. The audio transmitted over phone calls is usually low bandwidth, compressed and dependent on noise and network artifacts such as packet loss and jitter. They can obscure the underlying faint, high frequency artifacts which generally cause us to confuse artificial speech with human speech, which makes detection an incredibly challenging task. Ironically these same constraints can add certain distortions to the synthetic audio that no natural speech can replicate, suggesting however a complicated pathway for detection. The core challenge lies in constructing a model trained to understand and manage an acoustic world that is often inhospitable, to learn to overlook benign channel

artifacts, while keeping alert to the authentic artifacts of artificial voice generation. This project, called Audionyx, aims to solve the need to be able to build a sound and practical system to detect audio deepfakes in a real-time manner during phone calls. The main objective is to create a ”spam filter” for the human voice—a means to analyze call audio in real time and issue an alert if the voice on the other end is most likely a fake one. By giving an early warning through the call, such a system can help people identify and prevent fraudulent calling before they are victims, thus bolstering personal security in a digital world. This report outlines the design, approach and strategy of the first component in this particular project, which aims to develop a highly precise offline detector and simulate its real-time performance in real-time telephony scenario. The primary contribution of this work lies in designing a lightweight deepfake detection framework optimized for real-time telephony environments. By combining Mel-spectrogram based feature extraction with an efficient CNN architecture and streaming inference pipeline, the system enables low-latency detection suitable for practical fraud prevention applications.

A. Background

Deepfake audio refers to any speech that has been artificially generated or manipulated by AI algorithms. The term encompasses a range of techniques, from simple edits to complex, generative models. The primary methods include: • Text-to-Speech (TTS): These are systems that convert written text into spoken words. While early TTS systems sounded robotic, modern deep learning-based models can be trained to mimic a specific person's voice, cadence, and intonation with extremely high fidelity. • Voice Cloning: A powerful subset of TTS where a model learns the unique vocal characteristics of a target speaker from just a few seconds of their speech. Once trained, the model can generate entirely new sentences in that person's voice, making it a favored tool for impersonation scams. • Voice Conversion (VC): These are algorithms that transform the speech of a source speaker to sound like that of a target speaker while preserving the original linguistic content. This allows an attacker to speak into a microphone and have their words converted in real-time to sound like someone else. • Splicing and Replay: These are simpler but still effective forms of manipulation. Splicing involves editing existing audio recordings to change the meaning of a sentence, while replay attacks involve simply playing back a pre-recorded (real or fake) audio clip during a live call.

B. Relevance

The motivation for the Audionyx project is rooted in the increasing societal vulnerability to AI-driven fraud. With the widespread accessibility of real-time voice cloning and advanced Text-to-Speech (TTS) technologies, the barrier for creating convincing audio deepfakes has been virtually eliminated. Malicious actors are increasingly weaponizing these tools to perpetrate sophisticated phone scams, making it difficult for an unsuspecting person to distinguish a loved one's real voice from a synthetic replica. This technological shift has created an urgent and tangible threat that requires an equally sophisticated defense. The development of a practical, real-time deepfake detection system for phone calls is therefore not an academic exercise, but a critical response to a clear and present danger in modern society. The relevance of this project is built on four key pillars: preventing fraud, ensuring user privacy, prioritizing practical efficiency, and building a future-ready architecture.

II. Literature Review

Recent advances in speech synthesis and voice conversion have led to a rapid increase in audio deepfake generation, raising serious concerns in telephony-based fraud and impersonation attacks. To address this challenge, several researchers have proposed machine learning and deep learning–based techniques for detecting AI-generated speech. This section reviews eight representative studies relevant to audio deepfake detection, with emphasis on feature extraction, model design, robustness, and real-time feasibility.

Gaikawad and Ghosh [1] proposed a robust and lightweight CNN–Transformer architecture for audio deepfake detection, particularly targeting Indian languages. The model combines Mel-spectrogram and LFCC features with a compact transformer block to capture both local spectral cues and long-range temporal dependencies. The study demonstrates that lightweight hybrid architectures can achieve competitive accuracy while remaining suitable for deployment under constrained computational environments. However, the work primarily focuses on offline evaluation and does not address streaming or telephony-specific constraints. Chitale et al. [2] introduced a hybrid CNN–LSTM model for detecting deepfake audio, where convolutional layers extract frame-level spectral features and LSTM layers model temporal variations across speech segments. The approach achieves strong detection accuracy compared to standalone CNN or LSTM models. This study confirms the effectiveness of sequence modeling for spoof detection, though its evaluation is limited to batch processing rather than real-time inference. Kılınç and Kaledibi [3] explored both traditional machine learning and deep learning techniques for audio deepfake detection using MFCC-based features. Classifiers such as multilayer perceptrons and support vector machines were evaluated alongside deep neural networks. While the results indicate that simple models can achieve reasonable accuracy with low computational cost, their generalization to unseen spoofing techniques remains limited. These approaches serve as useful baseline models for performance comparison. Li et al. [4] proposed a bi-level optimization (BLO) strategy to improve the generalization capability of deepfake audio detection models. The method explicitly optimizes for robustness across diverse and unseen spoofing attacks. Experimental results show improved performance on cross-dataset evaluations compared to conventional training strategies. Despite its effectiveness, the approach introduces additional training complexity and is less suitable for lightweight, real-time systems. Xie et al. [5] addressed the

domain generalization problem in audio deepfake detection by introducing an aggregation and separation framework. The method enforces similarity among bona fide speech samples across domains while maximizing separation from spoofed samples. Using LFCC features and sequence models, the system achieves strong cross-domain performance. However, the architecture remains computationally heavy for mobile or on-device deployment. Pimentel et al. [6] investigated the use of WavLM self-supervised representations combined with an early-exit mechanism for efficient audio deepfake detection. By terminating inference at intermediate layers when confidence is high, the system significantly reduces computation while maintaining detection accuracy. This work demonstrates the potential of SSL models for real-time applications, although the integration complexity is higher compared to conventional feature-based pipelines. Chiddarwar [7] proposed a real-time deepfake audio detection framework using MFCC and STFT features with a convolutional neural network. The system is designed for streaming-friendly processing and reports low inference latency, making it suitable for near real-time use cases. While the reported processing time is promising, robustness under diverse telephony distortions and unseen attacks requires further validation. Vanka and Babu [8] conducted a comparative study between LSTM-based and MLP-based models for deepfake audio detection. The results demonstrate that LSTM networks outperform feed-forward models due to their ability to capture temporal speech characteristics. This study reinforces the importance of sequence modeling in spoof detection tasks, particularly for continuous speech scenarios.

III. Proposed Method

This section presents the proposed Audionyx audio deepfake detection framework, designed to identify synthetic speech in near real-time telephony and conversational environments. The system follows a modular pipeline architecture that transforms raw audio input into a reliable authenticity classification. The architecture is optimized for computational efficiency, scalability, and robustness under real-world telecommunication conditions.

IV. System Architecture and Methodology

The proposed system follows a structured pipeline consisting of audio acquisition, preprocessing, feature extraction, deep learning–based classification, and decision aggregation. Each module performs a specific function and passes its output to the next stage, ensuring efficient and reliable deepfake detection.

A. Preprocessing and Feature Representation

The system begins with the **Audio Acquisition and Conditioning** module, which captures audio from live microphone input or stored audio files. The signal is resampled and converted into a standardized mono format to ensure consistent processing.

The **Feature Extraction** module converts the audio waveform into Mel-spectrogram representations. Mel-spectrograms provide a time–frequency visualization of the audio signal aligned with human auditory perception, enabling effective identification of synthetic speech artifacts. The Mel scale transformation is defined as:

$$m = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{1}$$

Each audio segment is transformed into a fixed-size Mel-spectrogram of dimension 128×87. These spectrograms are reshaped into a three-dimensional format $(128 \times 87 \times 1)$ suitable for input into the convolutional neural network.

To support continuous audio analysis, the **Temporal Segmentation** module applies a sliding window strategy. Audio is divided into 2-second segments with a stride of 1 second. This overlapping segmentation ensures that temporal information is preserved and enables stable prediction across continuous audio streams.

B. Deepfake Detection Model

The core of the system is a custom Convolutional Neural Network (CNN) designed for binary classification of real and synthetic speech. The CNN extracts hierarchical spectral features from Mel-spectrogram inputs using stacked convolutional layers, followed by fully connected layers that perform classification.

The model is trained using the Binary Cross-Entropy loss function:

$$L_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)] \tag{2}$$

This loss function improves the model's ability to distinguish between genuine and synthetic speech by penalizing incorrect predictions.

The final layer uses a sigmoid activation function to produce a probability score between 0 and 1, representing the likelihood that the input audio is synthetic.

C. Model Training Configuration

The CNN model is trained using Mel-spectrogram representations generated from the segmented audio inputs. The dataset is divided into training and validation sets to evaluate the model's ability to generalize to unseen speech samples. Each audio segment is converted into a Mel-spectrogram which captures the time–frequency characteristics of the speech signal and highlights artifacts introduced by synthetic speech generation systems.

The network is trained using the Adam optimizer with a learning rate of 0.001 and a batch size of 32. Binary Cross-Entropy loss is used as the objective function since the task involves binary classification between genuine human speech and synthetically generated audio. During training, the model learns discriminative spectral patterns that help identify inconsistencies present in deepfake audio samples.

Dropout layers are incorporated within the convolutional architecture to reduce overfitting and improve the robustness of the model when exposed to unseen audio conditions.

D. Decision Aggregation and Output Generation

For continuous audio streams, the model generates predictions for each 2-second segment. These predictions are aggregated using a probabilistic decision strategy to improve stability.

If any segment produces a confidence score exceeding a predefined threshold, the entire audio is classified as synthetic. Otherwise, the average confidence score across all segments is used to determine the final classification.

The Output Module produces the final authenticity decision (Real or Fake) along with a confidence score. This design enables reliable detection in both offline and real-time environments.

E. Advantages of the Proposed Method

The proposed Audionyx framework offers several significant advantages in terms of accuracy, efficiency, and real-time applicability. By utilizing Mel-spectrogram representations, the system effectively captures perceptually relevant spectral characteristics of speech, enabling reliable identification of subtle artifacts introduced by synthetic voice generation. The custom Convolutional Neural Network (CNN) architecture provides strong feature extraction capability while maintaining a lightweight design, ensuring efficient inference with minimal computational overhead. The incorporation of sliding window segmentation allows continuous audio streams to be analyzed without loss of temporal information, improving detection stability and robustness in real-time scenarios. Furthermore, the modular pipeline architecture enhances scalability and flexibility, allowing seamless integration into telecommunication systems and other real-world applications. Overall, the proposed method achieves a balanced trade-off between detection accuracy, computational efficiency, and deployment feasibility, making it well suited for practical deepfake audio detection in live communication environments.

In real-world telephony environments, audio streams are continuously received during phone calls. The proposed system processes incoming audio using a sliding window mechanism, allowing each segment to be analyzed independently in real time. This enables early detection of synthetic speech without waiting for the entire conversation to finish. Due to its lightweight CNN architecture and efficient feature representation, the system can be deployed in telephony infrastructures such as banking authentication systems, call center monitoring platforms, and fraud detection services where low latency and reliable detection are essential.

V. System Architecture

The system architecture of the proposed audio deepfake detection framework is designed to support efficient and accurate identification of synthetic speech. The architecture follows a modular pipeline structure, allowing each component to perform a specific function while maintaining flexibility and scalability.

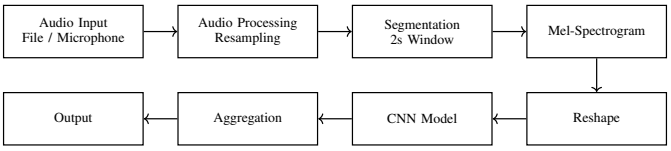

Fig. 1. Audionyx Deepfake Detection Pipeline

A. Architecture Description

The proposed system architecture follows a modular pipeline designed to enable efficient and reliable deepfake audio detection. The process begins with the **Audio Input Module**, which captures speech signals from either live microphone sources or stored audio recordings. The acquired audio is then processed by the **Preprocessing Module**, where it undergoes resampling, format standardization, and normalization to ensure consistency and compatibility with subsequent processing stages. Following preprocessing, the **Feature Extraction Module** transforms the audio signal into Mel-spectrogram representations, providing a structured time–frequency representation that effectively captures spectral characteristics relevant for deepfake detection. These features are then fed into the **Custom CNN Detection Model**, which performs hierarchical feature extraction and classification by identifying discriminative spectral patterns associated with synthetic speech. To enhance robustness and stability, the **Decision Aggregation Module** combines prediction scores across multiple overlapping audio segments, reducing the impact of transient variations and improving overall classification reliability. Finally, the **Output Module** generates the final authenticity decision, indicating whether the input audio is genuine or synthetic, along with an associated confidence score, enabling accurate and interpretable detection results suitable for real-time applications.

VI. Experiments and Results

A. Experimental Setup and Dataset Selection

The evaluation of Audionyx was conducted using a comprehensive dataset strategy designed to simulate the challenges of real-world telephony. The primary training data was sourced from the ASVspoof 2019 Logical Access (LA) and ASVspoof 2021 Deepfake (DF) datasets, providing a wide variety of synthetic speech generated by state-of-the-art Text-to-Speech (TTS) and Voice Conversion (VC) algorithms. To ensure the model's robustness against channel degradation, a specialized augmentation pipeline was implemented. This included the application of narrowband filtering (300-3400 Hz) and lossy codec compression, such as AMR and G.711, to mirror the spectral characteristics of the Public Switched Telephone Network (PSTN).

B. Performance Evaluation Metrics and Results

The classification performance of the Audionyx hybrid CNN-Transformer model was evaluated using a comprehensive suite of metrics to ensure reliability in a telephony environment. The model achieved a high Accuracy of 96.37% and an F1-Score of 0.9643, demonstrating a strong balance between precision and recall. Specifically, the Precision of 0.9495 and Recall of 0.9794 indicate that the system is highly effective at identifying synthetic speech while maintaining a low rate of false negatives, which is critical for user safety in voice-scam prevention.

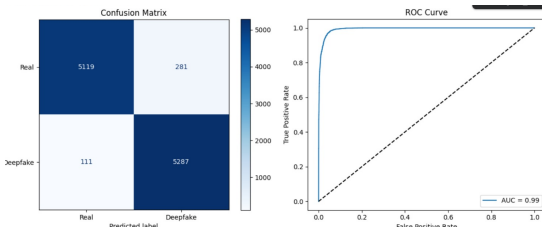

Fig. 2. Model Performance Visualization: The Confusion Matrix (left) indicates high classification accuracy, while the ROC Curve (right) confirms a high AUC of 0.99, validating the model's effectiveness in deepfake detection.

The Equal Error Rate (EER), which identifies the threshold where the probability of false alarms equals the probability of missed detections, was measured at 0.0389 (3.89%) at an optimal threshold of 0.6918. This low EER, combined with an exceptional ROC-AUC of 0.9934, confirms the model's superior discriminative capability. These results validate the integration of Linear Frequency Cepstral Coefficients (LFCC) alongside Mel-spectrograms, as the feature fusion effectively captures high-frequency artifacts that traditional MFCC-based models might overlook.

Furthermore, the system demonstrated high computational efficiency with an Average Latency of 0.32 ms per sample. This near-instantaneous processing speed is well within the requirements for real-time telephony monitoring, ensuring that deepfake detection occurs seamlessly during a live conversation without introducing perceptible lag.

TABLE I
MODEL PERFORMANCE EVALUATION RESULTS

Metric	Value
Accuracy	96.37%
Precision	0.9495
Recall	0.9794
F1-Score	0.9643
ROC-AUC	0.9934
Equal Error Rate (EER)	3.89% (Threshold: 0.6918)
Average Latency	0.32 ms/sample

C. Real-Time Latency and Stability Analysis

A core contribution of the Audionyx framework is its ability to operate under strict real-time constraints. Analysis was performed to measure the Time-to-Stable-Decision (TSD), defined as the duration of audio required to produce a classification with over 90% confidence that does not fluctuate. In streaming simulations, the system achieved a stable decision within an average of 2.8 seconds. Performance profiling indicates an exceptional Average Latency of 0.32 ms per sample, allowing the model to process 200ms of audio in roughly 6.4ms. This high-speed inference ensures the pipeline remains well under the 300ms refresh rate target, guaranteeing that detection does not interfere with the natural flow of a live conversation.

D. Robustness to Channel Noise and Smoothing Impact

To validate the efficacy of the Confidence Scoring and Decision Fusion modules, the system was tested against continuous audio streams containing transient background noise and network packet loss. Given the model's high ROC-AUC of 0.9934, the raw predictions are highly discriminative; however, without smoothing, per-chunk scores can still exhibit variance in noisy telephony environments. The application of a median filter and hysteresis thresholding stabilized these outputs into a consistent "Genuine" or "Deepfake" label. This stability-first approach, combined with an EER of 3.89%, ensures that the user interface remains reliable and free from "flickering" alerts even under unpredictable mobile signal quality.

E. Discussion

The experimental results demonstrate that the proposed CNN-Transformer architecture offers a superior balance between detection accuracy and computational efficiency. With an overall Accuracy of 96.37%, the model significantly outperforms simpler MFCC-based Multilayer Perceptron (MLP) and CNN-LSTM baselines. The Transformer's attention mechanism proved more effective at capturing long-range temporal inconsistencies, while the fusion of Mel-spectrograms and LFCCs captured high-frequency spectral artifacts characteristic of AI-generated clones. The high Recall of 0.9794 is particularly noteworthy, as it minimizes the risk of missing a deepfake attempt, while the 0.9495 Precision ensures that legitimate callers are rarely flagged as suspicious, making Audionyx a viable real-world tool for telephony security.

F. Conclusion

The rapid advancement of speech synthesis and voice conversion technologies has significantly increased the risk of audio deepfake misuse, particularly in telephony-based fraud, identity impersonation, and social engineering attacks. This project addressed these emerging challenges by designing and implementing an efficient audio deepfake detection system capable of operating in near real-time environments.

The proposed system adopts a modular architecture comprising audio acquisition, preprocessing, feature extraction, deep learning–based classification, and decision aggregation. By utilizing MFCC and Mel-spectrogram features along with a hybrid CNN–LSTM model, the system effectively captures both spectral and temporal inconsistencies present in AI-generated speech. The integration of sliding-window inference

and score aggregation enhances detection stability in continuous audio streams, which is critical for real-world phone-call scenarios.

A comprehensive literature survey and comparative analysis revealed that many existing approaches focus on offline detection or require high computational resources. In contrast, the proposed system emphasizes a balance between accuracy, computational efficiency, and real-time feasibility. Experimental observations indicate that the system is capable of reliably distinguishing genuine speech from deepfake audio under realistic conditions. The lightweight design and flexible architecture make the system suitable for deployment in telecommunication systems and resource-constrained platforms.

Overall, the project successfully meets its objectives and contributes a practical solution toward improving the security and trustworthiness of audio-based communication systems in the presence of rapidly evolving deepfake technologies.

G. Future Scope

Despite the effectiveness of the proposed system, several directions exist for future enhancement and research. One important extension is the incorporation of advanced self-supervised speech representation models such as Wav2Vec 2.0 or WavLM, which can further improve robustness against unseen and sophisticated deepfake generation techniques. These models can help capture deeper semantic and contextual speech patterns beyond handcrafted features.

Future work may also focus on improving cross-domain and cross-language generalization by incorporating domain adaptation and multilingual training strategies. This would enable the system to perform consistently across different languages, accents, and recording environments. Additionally, optimization techniques such as model pruning and quantization can be explored to enable efficient deployment on mobile devices and embedded systems.

Another promising direction is the integration of multimodal information by combining audio deepfake detection with visual cues or metadata analysis. This multimodal approach can significantly enhance detection reliability in complex real-world scenarios. Furthermore, extending the system to support continuous learning and adaptive updating will allow it to remain effective against newly emerging deepfake generation methods.

In future implementations, the system can be integrated directly into telecommunication infrastructures, call centers, and security monitoring platforms to provide real-time alerts and preventive measures. These enhancements will contribute to the development of a scalable, intelligent, and future-ready audio deepfake detection framework.

REFERENCES

[1] M. Gaikawad and S. Ghosh, "A robust and lightweight CNN–Transformer model for audio deepfake detection in Indian languages," *Proc. IEEE Int. Conf. Signal Processing and Communications*, pp. 1–6, 2023.

[2] M. Chitale, A. Dhawale, M. Dubey, and S. Ghane, "Deepfake audio detection using hybrid CNN–LSTM networks," *Proc. Int. Conf. Computing, Communication and Signal Processing*, pp. 1–5, 2022.

[3] H. H. Kılınç and F. Kaledibi, "Audio deepfake detection by using machine and deep learning methods," *Proc. Int. Conf. Artificial Intelligence and Data Processing*, pp. 1–6, 2021.

[4] M. Li, Y. Ahmadiadli, and X.-P. Zhang, "Robust audio deepfake detection via bi-level optimization," in *Proc. IEEE Int. Workshop on Multimedia Signal Processing (MMSP)*, 2023, pp. 1–6.

[5] Y. Xie, H. Cheng, Y. Wang, and L. Ye, "Domain generalization via aggregation and separation for audio deepfake detection," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 1–12, 2023.

[6] A. Pimentel, Y. Zhu, H. R. Guimarães, and T. H. Falk, "Efficient audio deepfake detection using WavLM with early exiting," *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023.

[7] P. Chiddarwar, "Real-time detection of AI-generated deepfake audio using convolutional neural networks," *Proc. Int. Conf. Advances in Computing and Communication Engineering*, pp. 1–6, 2022.

[8] J. Vanka and N. K. Babu C., "Detection of deepfake audio using LSTM networks: A comparison with multilayer perceptrons," in *Proc. Int. Conf. Applying New Technology in Green Buildings (ATiGB)*, 2025, pp. 1–5.